Segment-based Acoustic Models for Continuous Speech Recognition

Progress Report: 1 January 94 – 31 March 94

submitted to
Office of Naval Research
and
Advanced Research Projects Administration
11 May 1994

by
Boston University
Boston, Massachusetts 02215

Principal Investigators

Dr. Mari Ostendorf
Associate Professor of ECS Engineering, Boston University
Telephone: (617) 353-5430

Dr. J. Robin Rohlicek
Division Scientist, BBN Inc.
Telephone: (617) 873-3894

Administrative Contact

Maureen Rogers, Awards Manager
Office of Sponsored Programs
Telephone: (617) 353-4365

Executive Summary

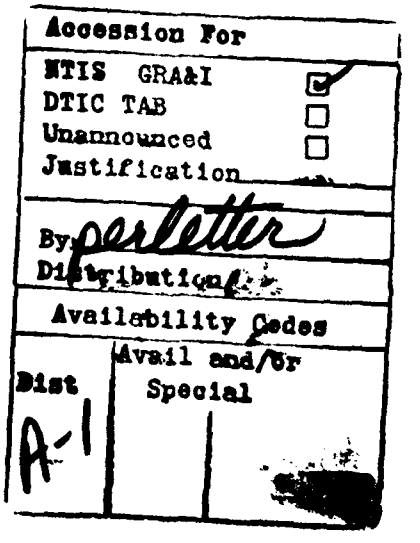
This research aims to develop new and more accurate stochastic models for speaker-independent continuous speech recognition by extending previous work in segment-based modeling and by introducing a new hierarchical approach to representing intra-utterance statistical dependencies. These techniques, which have high computational costs because of the large search space associated with higher order models, are made feasible through rescoring a set of HMM-generated N-best sentence hypotheses. We expect these different modeling techniques to result in improved recognition performance over that achieved by current systems, which handle only frame-based observations and assume that these observations are independent given an underlying state sequence.

In the past quarter, our focus has been on developing the theory and initial implementation behind high level models and search algorithms to accommodate these models. These efforts and other accomplishments of this project include:

- further investigation of different variations of mixture distributions, looking in particular at robust estimation and initialization issues;
- development of a new approach to continuous density parameter adaptation;
- development of training algorithms for a discrete distribution intra-utterance correlation model and two methods for applying the model in multi-pass recognition scoring;
- conducted further experiments with the sentence-level mixture language model, demonstrating performance improvements on the 5k vocabulary WSJ hub test set;
- implemented much of the software needed for lattice rescoring with the SSM;
- modified our distribution clustering algorithm to allow general splits; and
- began exploring auditory signal processing algorithms on the TIMIT phone recognition task.

We also continued to make minor improvements to our baseline recognition system. Our current best performance figures with the BU acoustic and language model are 7.0% and 5.7% error rates on the 5k WSJ development and evaluation test sets, respectively. These numbers improve to 6.3% and 5.3% error rates when we also include the BBN HMM and SNN acoustic scores in reranking the sentence hypotheses.

Contents



Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1994 – 31 March 1994

1 Productivity Measures

- Refereed papers submitted but not yet published: 0
- Refereed papers published: 0
- Unrefereed reports and articles: 1
- Books or parts thereof submitted but not yet published: 0
- Books or parts thereof published: 0
- Patents filed but not yet granted: 0
- Patents granted (include software copyrights): 0
- Invited presentations: 1
- Contributed presentations: 1
- Honors received: none
- Prizes or awards received: none
- Promotions obtained: none
- Graduate students supported $\geq$ 25% of full time: 5
- Post-docs supported $\geq$ 25% of full time: 0
- Minorities supported: 2 women

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1994 – 31 March 1994

2 Summary of Technical Progress

Introduction and Background

In this work, we are interested in the problem of large vocabulary, speaker-independent continuous speech recognition, and primarily in the acoustic modeling component of this problem. In developing acoustic models for speech recognition, we have conflicting goals. On one hand, the models should be robust to inter- and intra-speaker variability, to the use of a different vocabulary in recognition than in training, and to the effects of moderately noisy environments. In order to accomplish this, we need to model gross features and global trends. On the other hand, the models must be sensitive and detailed enough to detect fine acoustic differences between similar words in a large vocabulary task. To answer these opposing demands requires improvements in acoustic modeling at several levels: the frame level (e.g. signal processing), the phoneme level (e.g. modeling feature dynamics), and the utterance level (e.g. defining a structural context for representing the intra-utterance dependence across phonemes). This project addresses the problem of acoustic modeling, specifically focusing on modeling at the segment level and above. The research strategy includes three main thrusts. First, phone-level acoustic modeling is based on the stochastic segment model (SSM) [1, 2], and in this area our main efforts involve developing new techniques for robust context modeling, mechanisms for effectively incorporating segmental features, and models of within-segment dependence of frame-based features. Second, high-level models are being explored in order to capture speaker-dependent and session-dependent effects within the context of a speaker-independent model. In particular, we are investigating hierarchical structures for representing the intra-utterance dependency of phonetic models, and more recently language models for representing topic dependency and language dynamics, recognizing that higher-order models of correlation can extend to the language domain as well as the acoustic domain. Lastly, speech recognition is implemented under the N-best rescoring paradigm [3], in which the BBN Byblos system is used to constrain the SSM search space by providing the top N sentence hypotheses. This paradigm facilitates research on high-order models through reducing development costs, and provides a modular framework for technology transfer that has enabled us to advance state-of-the-art recognition performance through collaboration with BBN.

Summary of Technical Results

In the first half of this project, we have focused on improving the performance of the basic segment word recognition system and porting the system to the Wall Street Journal task domain. In brief, the accomplishments of that period included: improvements to the N-Best rescoring weight estimation algorithm; investigation of different mechanisms for improving the baseline acoustic model, including distribution clustering [4], mixture modeling at different time scales [5, 6], theoretically consistent models based on context-dependent posterior distributions, automatic distribution mapping estimation, and hierarchical models of intra-utterance phoneme dependence; implementation of baseline n-gram and sentence-level mixture language models; and improvements to the SSM baseline system aimed at improving performance on the ARPA WSJ task and participation in the benchmark tests.

The research efforts during this quarter, supported in part by an ONR AASERT award, have primarily involved theoretical and initial software development for models of high-order correlation and search algorithms to accommodate these models. These efforts and other research developments are summarized below.

Mixture distributions in the SSM. We have been attempting to improve the performance of the untied mixture system, both for the independent-frame model and the segmental mixture model, investigating issues related to robust parameter estimation. One issue that arises in training mixtures is that care must be taken to ensure that covariance estimates are robust. The EM algorithm for training mixtures, like all maximum likelihood (ML) methods, suffers from the difficulty that the training likelihood can be increased without bound if a Gaussian mean is centered on one of the training observations and its covariance is made to approach a singularity. ML training algorithms accordingly tend to naturally converge to such degenerate solutions if not corrected in some way. We investigated a number of variance "clipping" schemes for diagonal covariance Gaussians, patterned after the approach given in [7]. In these experiments, the degree of clipping was varied according to the amount of training for the Gaussian. We also looked at clipping the variance based on a percentage of two different values – either a histogram of the estimated variances or an estimate of the overall feature variance. Unfortunately, these methods led to only very small improvements over the previous method of clipping with a single threshold, regardless of training. Investigating this further, we found that other researchers have observed similar results [8]. We have also been investigating alternate initialization procedures, including one that starts from a single Gaussian per mixture and successively increases the number of components by "splitting" the Gaussian with the largest likelihood. This approach has so far produced inferior performance to our previous approach of initializing from non-mixture SSM models.

Adaptation of continuous distributions. It is well known that speech recognition performance can be improved by matching the training data to the test conditions, whether it be to address channel/environment effects or match the recognizer to a particular speaker. In most cases, however there is not enough training data available to handle a particular speaker/environment condition, and automatic adaptation techniques are used to tune a general model to a particular condition. In particular, we are interested in unsupervised, incremental adaptation of Gaussian density parameters, which is more difficult than adaptation of tied mixture parameters (the approach currently used by most sites) because of the large number of parameters to adapt. Our approach is to combine the advantages of a Bayesian framework [7, 9] with the robust properties of vector field smoothing [10] to adapt a large number of continuous densities. To handle the problem of adapting a large number of parameters, we define a two-level model, that includes a large set of distributions used for recognition that we would like to adapt and a smaller set of distributions that represent a coarser but more frequently observed space. This two level model is built using divisive, maximum likelihood distribution clustering, which gives a tree with detailed models at its leaves (to be adapted) and coarser models defined by a pruned version of the tree (for which adaptation statistics are computed). Like vector field smoothing, we use a weight function for determining the relative contribution of adaptation vectors; but in our case the weight is based on a measure of the mutual information between the coarse and fine distributions, rather than Euclidean distance as used in vector field smoothing. An advantage of this framework is that the coarse models can easily be chosen to reflect different degrees of detail and thus tuned to the amount of adaptation data available, i.e. a small set of adaptation distributions is appropriate for a small set of adaptation data. The adaptation algorithm is currently being implemented for uni-modal, full-covariance Gaussian mean adaptation.

Mixture language modeling. One of the important questions in language modeling today is how to effectively represent the long-term structure of language, i.e. how to capture dependence over longer sequences of words than can be modeled with a simple n-gram. To address this problem, we have developed a sentence-level mixture language model (LM) that represents the topic-dependent structure of language with separate n-gram language model mixture components determined using automatic clustering. We have continued the experimental work that we started in the previous quarter, but in our recent efforts we were able to use additional language model training data obtained from BBN. We found a small reduction in the perplexity due to the mixture language models using 5 and 8 component mixtures to model the 5k vocabulary WSJ task. More importantly, we obtained a 6% reduction in the recognition error using the 5-component mixture models as compared to the standard trigram models. A paper reporting these results was presented at the March 1994 ARPA Workshop on Human Language Technology [11]. In all the experiments so far where we sought to demonstrate feasibility of the model, clustering and estimation algorithm approximations were used to save computation. Current efforts are aimed at implementing more theoretically motivated algorithms, such as a similarity measure using inverse document frequencies and full

Expectation-Maximization (EM) training, which we expect will result in a small improvement in the models.

Intra-utterance phoneme dependence modeling. We further developed the theoretical framework for a hierarchical model of dependence for a set of discrete random variables, which we have begun investigating as a model of intra-utterance phoneme dependence. We use a dependence tree [12] to represent the correlation among random variables, using a tree structure (designed automatically) with Markov assumptions along the branches of the tree. The dependence tree gives us an approximation to the joint distribution of the set of random variables with a much smaller number of parameters than a full joint model. We can apply this model to speech recognition, by defining a fixed-length vector (one dimension per phoneme) for each utterance and letting each element of the vector take on a discrete value (e.g. vector quantization index) based on all of the versions of the phone in the utterance that corresponds to that element of the vector. Any phone that did not appear in the utterance is considered to be "unobserved". The dependence tree model represents a distribution for this vector that can be used in recognition in at least two ways. First, it can provide a "score" of the consistency of the phones in the utterance and incorporated as an additional knowledge source in N-best rescoring. Second, the dependence tree can serve as a prior on phone distribution parameters, which would be used in a Bayesian approach parameter adaptation or even directly in recognition scoring. Since we have unobserved data, the Expectation-Maximization (EM) algorithm is required for parameter estimation. Although we eventually plan to use the EM algorithm, our initial implementation of the training algorithm has been based on a simple iterative algorithm that substitutes the most likely observation for a missing component. An initial model has been estimated using simple cepstral features on a subset of the TIMIT corpus, which gives an increase in training data likelihood as expected. Our next step is to implement EM training and build a model on the full corpus, before turning to the implementation of the algorithm for recognition.

Lattice search algorithms for multi-pass recognition scoring. In the previous quarter, we developed a local search algorithm for lattice rescoring, motivated by the good performance of our recognition system on the WSJ (which suggest that we could improve performance by considering more hypotheses) and the need for a framework that can accommodate sentence-level acoustic and language models. We plan to implement both optimal and local lattice search algorithms, so as to assess the performance/speed trade-offs experimentally, and in this quarter we initiated software development for extracting the lattice and dynamic programming search. The effort involves both developing software at BU for performing the search using the SSM and modifying the BBN decoder so that it produces word lattices. So far, an SSM word lattice rescoring algorithm has been implemented that handles cross-word triphone models and trigram language models. The algorithm uses hypothesized phone boundary windows to reduce computation, where the windows

can be determined from phone end times provided by the previous HMM scoring pass or by a prediction based on left-context phone duration means. Experiments are underway to verify that performance is comparable to N-best rescoring and to assess the impact of using predicted phone end times. In addition, we modified the BBN decoder to produce lattices from the N-best traceback structure, but further modifications are needed to produce silence information (i.e. the number of frames of silence at the end of each word) and optionally phone segmentations for each word. [This work was supported by the ONR AASERT award.]

Distribution clustering for model size reduction. In previous work, we developed a distribution clustering algorithm to specify regions of parameter tying, effectively reducing model size without sacrificing performance. The algorithm was based on divisive clustering, similar to that used in HMMs for state-dependent clustering, but with a maximum likelihood criterion specified according to the specific parameters to be tied [4]. One of the limitations of any tree-based clustering algorithm is that successive divisions of the training data can lead to data fragmentation and less effective use of context information. This is potentially a greater problem for our system in that we implemented only simple questions of the form "is the ith feature in set A?" To partially overcome this problem, we implemented complex (multi-feature) questions using a version of the "pylon" question design algorithm [13]. A pylon contains a sequence of binary questions tied so that at each stage there are only two subdivisions of the data. In tree design, we first choose a question and node to split as in normal growing of the tree. Then a pylon growing algorithm is used, which successively chooses the best question to split and then the best merge out of two of the three data subsets, where both splits and merges are evaluated according to our maximum likelihood criterion. This procedure is continued until the stopping criterion is reached (maximum number of questions asked in the pylon or no question further splits the nodes) and the remaining two nodes are taken as the child nodes produced due to the split by the original question at the splitting node. In preliminary results, we find that when we use pylons to further reduce model size by about 50%, we observe slightly worse performance compared to the normal clustering. More experiments are needed to determine whether performance gains are achieved when model size is the same as for our standard clustering.

Auditory-based signal processing. We have begun an effort to assess the use of different signal processing algorithms in speech recognition, motivated by capabilities of the human auditory system. Our goal is to demonstrate improved recognition performance for clean signals, not just signals in noise as is typically done with auditory models). To reduce the scope of this project, we are assessing phone recognition performance on the TIMIT task using HTK (Hidden Markov Model Toolkit). With the help of Steve Young from Cambridge University, we have duplicated a system based on cepstral parameters that gives one of the best reported results on this task [14], and we are currently assessing the performance of the Seneff auditory model (available through the

Frontiers in Speech Recognition Workshop CDROM) using a similar strategy for HMM topology building. Initial results are not encouraging, but we hope with further experiments to establish baseline performance at least close to that for cepstra, and then to consider alternative auditory models. [This work was supported by the ONR AASERT award.]

Future Goals

Based on the results of the past year and our original goals for the project, we have set the following goals for the next six months: (1) continue implementation of the lattice search algorithm and assess performance/speed trade-offs; (2) further develop the hierarchical model formalism and assess the trade-offs between linear and non-linear models of dependence; (3) extend the language modeling work include a dynamic component and to handle new vocabulary items; and (4) investigate unsupervised adaptation in the WSJ task domain.

References

[1] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. Acoustics Speech and Signal Processing, Dec. 1989.*

[2] S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *IEEE Int. Conf. Acoust., Speech, Signal Processing,* pages 127–130, New York, New York, April 1988.

[3] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proc. of the DARPA Workshop on Speech and Natural Language,* pp. 83-87, February 1991.

[4] A. Kannan, M. Ostendorf and J. R. Rohlicek, "Maximum Likelihood Clustering of Gaussians for Speech Recognition," *IEEE Transactions on Speech and Audio Processing,* to appear.

[5] A. Kannan and M. Ostendorf, "A Comparison of Trajectory and Mixture Modeling in Segment-based Word Recognition," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Processing,* Vol. II, pp. 327-330, April 1993.

[6] O. Kimball and M. Ostendorf, "On the Use of Tied Mixture Distributions," *Proceedings of the ARPA Workshop on Human Language Technology,* pp. 102-107, 1993.

[7] C.H. Lee, C.H. Lin, and B.H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans on Signal Processing,* Vol. 39, No. 4, April 1991, pp. 806–814.

[8] J. L. Gauvain, personal communication.

[9] B. Necioglu, M. Ostendorf and J. R. Rohlicek, "A Bayesian Approach to Speaker Adaptation of the Stochastic Segment Model," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Processing,* March 1992, Vol. I, pp. 437-440.

[10] K. Ohkura, M. Sugiyama and S. Sagayama, "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," *Proc. of the Inter. Conf. on Spoken Language Processing,* Vol. 1, pp. 369-372.

[11] R. Iyer, M. Ostendorf and J. R. Rohlicek, "Language Modeling with Sentence-Level Mixtures," *Proceedings of the 1994 ARPA Workshop on Human Language Technology,* March 1994.

[12] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory,* Vol. IT-14, No. 3, May 1968, pp. 462-467.

[13] L. R. Bahl, P. F. Brown, P. V. deSouza and R. L. Mercer, "A Tree-Based Statistical Language Model for Natural Language Speech Recognition," *IEEE Trans. on Acoust., Speech, and Signal Processing,* Vol. 37, No. 7, pp. 1001-1008, 1989.

[14] S. J. Young and P. Woodland, "The Use of State Tying in Continuous Speech Recognition," *Proc. Eurospeech,* pp. 2203-2207, 1993.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1994 - 31 March 1994

3 Publications and Presentations

During this reporting period, we published one conference paper (abstracts are reviewed), and gave an additional conference presentation, as well as one invited talk associated with this project, as itemized below.

"Language Modeling with Sentence-Level Mixtures," R. Iyer, M. Ostendorf and J. R. Rohlicek, *Proceedings of the 1994 ARPA Workshop on Human Language Technology*, March 1994.

"Stochastic Segment Modeling for Continuous Speech Recognition," M. Ostendorf *et al.*, presented at the March 1994 ARPA Workshop on Spoken Language Technology.

"A Unified View of Stochastic Modeling for Speech Recognition", M. Ostendorf, invited talk at ICSI, Berkeley, CA, January 1994

In addition, one Boston University M.S. thesis proposal was successfully defended by Fred Richardson, entitled "Lattice-Based Search Strategies for the Stochastic Segment Model."

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1994 - 31 March 1994

4 Transitions and DoD Interactions

This grant includes a subcontract to BBN, and the research results and software is available to them. Thus far, we have collaborated with BBN by combining the Byblos system with the SSM in N-Best sentence rescoring to obtain improved recognition performance, and we have provided BBN with papers and technical reports to facilitate sharing of algorithmic improvements. On their part, BBN has been very helpful to us in our WSJ porting efforts, providing us with WSJ data and consulting on format changes. We have also begun an effort to collaborate more closely in lattice rescoring, and expect that Boston University student Fred Richardson will implement software libraries that will be shared by both sites.

The recognition system that has been developed under the support of this grant and of a joint NSF-ARPA grant (NSF # IRI-8902124) is currently being used for automatically obtaining good quality phonetic alignments for a corpus of radio news speech under development at Boston University. The alignment effort is supported by the Linguistic Data Consortium, through a grant that allowed us to add cross-word phonological rules to the segmentation software.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1994 - 31 March 1994

5 Software and Hardware Prototypes

Our research has required the development and refinement of software systems for parameter estimation and recognition search, which are implemented in C or C++ and run on Sun Sparc workstations. No commercialization is planned at this time.